TWO-ARMED BANDITS WITH A GOAL;
II: DEPENDENT ARMS

by

Donald A. Berry*

and

Bert Fristedt**

University of Minnesota

Technical Report No. 356

September, 1979

*This author's research sponsored by the NSF under Grant No. 78-02694.

**This author's research sponsored by the NSF under Grant No. MCS 78-01168 A01.

Two-armed bandits with a goal; II: Dependent arms

by

Donald A. Berry and Bert Fristedt

ABSTRACT

One of two random variables, X and Y, can be selected at each of a possibly infinite number of stages. Depending on the outcome, one's fortune is either increased or decreased by one. The probability of increase may not be known for either X or Y. The objective is to increase one's fortune to G before it decreases to g, for some integral g and G; either may be infinite.

In Part I (Berry and Fristedt 1979), the distribution of X is unknown and that of Y is known. In the current part, it is known that either X or Y has probability α of increasing the current fortune by one and the other has probability β of increasing the fortune by one, where α and β are known, but which goes with X is not known. We show that optimal strategies exist in general and find all optimal schemes when $\alpha = 0$ and when $\alpha + \beta = 1$. In both cases myopic strategies are shown to be optimal. A counterexample is used to show that myopic strategies, while intuitively very appealing, are not optimal for general (α, β).

Key words and phrases: Achieving a goal, two-armed bandits, how to gamble if you must, gambler's ruin, sequential decisions, Bayesian decision making, sequential medical treatments, stochastic control, optimal dynamic designs, myopic strategies.

1. Introduction. In this paper we consider a variant of the problem presented by Berry and Fristedt (1979). Briefly, a decision maker is able to choose one of two experiments to perform (or one of two "arms" $\mathfrak{L}$ and $\mathfrak{R}$ to pull) at each of an unlimited number of stages. His objective is to convert his current fortune, k, into G before it becomes g, where $g < k < G$. Each pull of an arm increases or decreases his fortune by 1; choosing $\mathfrak{L}$ increases k with probability λ and choosing $\mathfrak{R}$ increases k with probability ρ. In Part I (Fristedt and Berry 1979), we consider the case in which λ is known.

In the current paper, λ and ρ are both unknown but are related in a very special way: one arm has success probability α and the other has success probability β for given constants α and β, but it is not known which goes with which arm. With no further loss of generality we take $\beta \geq \alpha$, and let the initial probability that $(\lambda, \rho) = (\alpha, \beta)$ be r. Define $\underset{\sim}{Q}_{\alpha,\beta}$ to be the class of such two-point distributions:

$$F \in \underset{\sim}{Q}_{\alpha,\beta} \Leftrightarrow \text{supp } F \subset \{(\alpha,\beta), (\beta,\alpha)\}.$$

Let

$$\underset{\sim}{Q} = \bigcup_{0\leq\alpha\leq\beta\leq 1} \underset{\sim}{Q}_{\alpha,\beta} .$$

We now present adapted versions of the notation and terminology that was developed in (Berry and Fristedt 1979). The current discussion and theorems can be read independently, but to follow the proofs the reader is advised to refer to Sections 1 and 2 of that paper.

For an arbitrary strategy, let r_n denote the probability that $(\lambda, \rho) = (\alpha, \beta)$ conditioned on the information from the first n pulls (that is, at the $(n+1)$st stage) and let k_n denote the fortune at that time. Let $\sigma_{\mathfrak{A}} r$ denote the new probability that $(\lambda, \rho) = (\alpha, \beta)$ after a success has been obtained with arm $\mathfrak{A}$ and $\varphi_{\mathfrak{A}} r$ denote the conditional probability that $(\lambda, \rho) = (\alpha, \beta)$ given a failure with arm $\mathfrak{A}$ ($= \mathfrak{L}$ or $\mathfrak{R}$). For example, if $\mathfrak{R}$ is pulled initially, then

$$(r_1, k_1) = (\sigma_{\mathfrak{R}} r, k+1) \text{ with prob. } r\beta + (1-r)\alpha$$

$$= (\varphi_{\mathfrak{R}} r, k-1) \text{ with prob. } r(1-\beta) + (1-r)(1-\alpha),$$

where

$$\sigma_{\mathfrak{R}} r = \frac{r\beta}{r\beta + (1-r)\alpha} \quad ,$$

$$\varphi_{\mathfrak{R}} r = \frac{r(1-\beta)}{r(1-\beta) + (1-r)(1-\alpha)} \quad .$$

For fixed α and β, these notations and extensions such as $\sigma_{\mathfrak{L}}^2 \varphi_{\mathfrak{R}}^3 \sigma_{\mathfrak{R}} r$ are unambiguous except that certain such expressions may not be defined in some trivial cases such as $r = \beta = 1$. Two easily obtained, but useful, relations are:

(1.1) $\quad \sigma_{\mathfrak{L}} \sigma_{\mathfrak{R}} r = r \ ,$

(1.2) $\quad \varphi_{\mathfrak{L}} \varphi_{\mathfrak{R}} r = r \ ,$

valid when the left sides are defined.

For fixed α, β, g, and G, a <u>scheme</u> ψ is a function $\psi(k, r)$ which

assigns either of the symbols $\mathcal{R}$ or $\mathcal{L}$ to each $k \in (g, G)$ and $r \in [0, 1]$. The strategy induced by a scheme Ψ requires a pull of $\Psi(k_n, r_n)$ at the $(n + 1)$st stage, for $n = 0, 1, \ldots$. We shall rely on Theorems 2.1 and 2.2 of Berry and Fristedt (1979) which give conditions for a scheme to be optimal. These theorems apply in the current setting with obvious modifications.

Optimal strategies for the classical two-armed bandit (in which the expected number of successes is to be maximized) with initial distribution in $\underset{\sim}{Q}$ was found by Feldman (1962) -- see also (DeGroot 1970, Section 14.7) -- , and generalized in different directions by Fabius and van Zwet (1970) and Berry (1972). Feldman (1962) showed that an optimal scheme is to always pull the arm which has the higher probability of being better; $\mathcal{L}$ if $r < .5$, $\mathcal{R}$ if $r > .5$, and either -- say $\mathcal{R}$ for definiteness -- if $r = .5$. This scheme, call it Ψ_0, is sometimes called "myopic." Kelley (1974) has completely characterized two-point distributions in the square for which myopic schemes are optimal in the classical two-armed bandit.

Myopic schemes are also compelling in the current problem. Suppose $r < .5$; not only is the probability of success with $\mathcal{L}$ greater than with $\mathcal{R}$ (that is, $E\lambda > E\rho$), but also $Em(\lambda) > Em(\rho)$ for any increasing function m on $\{\alpha,\beta\}$. In particular, for any s the probability of s immediate successes with $\mathcal{L}$ is greater than with $\mathcal{R}$: $E\lambda^s > E\rho^s$. Incredibly, as Example 3.1 shows, Ψ_0 is not always optimal! The example also suggests how difficult it is to find an optimal strategy

in general. Though the case considered is quite special, we cannot actually find an optimal strategy, we merely find one that is better than the one induced by Ψ_0. The prospect of finding optimal strategies, or or even partially characterizing optimal schemes, for general measures on $(\rho,\lambda) \in [0, 1] \times [0, 1]$ seems quite remote.

We <u>are</u> able to completely characterize optimal schemes for two special subsets of $\underset{\sim}{Q}$. In Section 4 we resolve the case $\alpha = 0$ and in Section 5 the case $\alpha + \beta = 1$. In both cases we find that myopic schemes are optimal, but when $\alpha + \beta = 1$ there are many optimal schemes that are not myopic.

The case $\beta = 1$ is at the opposite extreme from $\alpha = 0$. However, since a single failure (on either arm) gives complete information, $g + 1$ is the only fortune of any mathematical interest. The interested reader can easily verify that Ψ_0 is optimal in this case as well, but there are a great many other optimal schemes (unless $G - g = 2$).

In Section 6 we draw an analogy with certain cases in which λ and ρ have independent two-point distributions. In the next section we show that optimal schemes exist for $\underset{\sim}{Q}$.

2. Existence of Optimal Schemes. Let

$$U(k, r) = \sup_{\tau} P_{\tau}^{(k, r)} \quad (k_n \to G)$$

where the supremum is taken over all possible strategies τ for beginning at (k, r). (In case $G < \infty$, $k_n \to G$ means that $k_n = G$ for sufficiently large n.) Schemes as well as strategies can be used as subscripts. A scheme Ψ is conserving if

$$(2.1) \qquad E_{\Psi}^{(k,r)} U(k_1, r_1) = U(k, r) \ .$$

Optimal schemes are obviously conserving. To prove the existence of optimal schemes for $\underset{\sim}{Q}$ we shall use Theorem 2.1 of our companion paper (Berry and Fristedt 1979), which gives a sufficient condition for a conserving scheme to be optimal.

THEOREM 2.1. For every g and G there exists an optimal scheme for $\underset{\sim}{Q}$.

PROOF. There exists a conserving scheme for $\underset{\sim}{Q} - U\underset{\sim}{Q}_{\alpha,1}$. When $g > -\infty$ Theorem 2.1 of Berry and Fristedt (1979), applies to show that such a scheme is optimal for $\underset{\sim}{Q} - U\underset{\sim}{Q}_{\alpha,1}$. If $g = -\infty$ and $\beta < .5$ then the same theorem again applies to show that a conserving scheme is optimal for $Q_{\alpha,\beta}$.

Suppose $g = -\infty$ and $\beta \geq .5$. Let Ψ_0 be the myopic scheme defined in Section 1; that is,

$$(2.2) \qquad \Psi_0(k, r) = \mathfrak{R} \text{ if } r \geq .5,$$
$$= \mathfrak{L} \text{ if } r < .5.$$

Assume that $\alpha \neq \beta$ and that G is not reached when following Ψ_0. By the strong law of large numbers, the arm whose success probability is α will be pulled only finitely often under Ψ_0. Therefore, Ψ_0 is optimal for each $Q_{\alpha,\beta}$ with $\beta \geq .5$.

It is also clear that Ψ_0 is an optimal scheme for $\cup \underset{\sim}{Q}_{\alpha,1}$ in case $g > -\infty$. ∎

In the next section we show that Ψ_0 is not optimal for $\underset{\sim}{Q}$, at least in case $G - g = 3$.

3. Counter-example. We now present what to us is one of the most counter-intuitive examples we have seen in bandit problems. We shall see that there are distributions in $\underset{\sim}{Q}$ for which it is not optimal to pull the arm which has the higher probability of being the β-arm. This is in contrast to Feldman's (1962) result for classical bandits.

EXAMPLE 3.1. Let $g = 0$ and $G = 3$. Since the function $x \mapsto x^3(1-x)$ is increasing on $[0, .75]$ and decreasing on $[.75, 1]$ we can choose $0 < \alpha < \beta < 1$ so that

$$\alpha^3(1 - \alpha) = \beta^3(1 - \beta) \quad . \tag{3.1}$$

We shall need the inequality $\alpha + \beta > 1$, a consequence of the fact that $x^3(1-x) < (1-x)^3x$ for $x \in (0, .5)$. For Ψ_0 defined by (2.2) and a particular (α, β) satisfying (3.1), we shall show that τ_0, the strategy induced by Ψ_0 for starting at $(2, (1-\alpha)/(2-\alpha-\beta))$, is not optimal. Let

$$W_0(k, r) = P_{\Psi_0}^{(k,r)}(k_n \to 3) \quad .$$

The fact that the process steps back and forth between $k = 1$ and 2 when is does not terminate at 0 or 3 facilitates our analysis.

When using Ψ_0 and starting at $k = 1$ with $r = .5$, $\mathfrak{R}$ is pulled twice, then $\mathfrak{L}$ twice, then $\mathfrak{R}$ twice, etc., until, of course, 0 or 3 is reached. For, by easy calculations,

$$\sigma_{\mathfrak{R}} r = \beta/(\alpha + \beta) > .5 \quad ,$$

$$\varphi_{\mathcal{R}}\sigma_{\mathcal{R}}r = \beta(1-\beta)/[\beta(1-\beta) + \alpha(1-\alpha)] < .5 ,$$

$$\sigma_{\mathcal{L}}\varphi_{\mathcal{R}}\sigma_{\mathcal{R}}r = (1-\beta)/[1-\beta+1-\alpha] < .5 ,$$

$$\varphi_{\mathcal{L}}\sigma_{\mathcal{L}}\varphi_{\mathcal{R}}\sigma_{\mathcal{R}}r = .5$$

-- the last fact follows as well from (1.1) and (1.2). Hence

$$W_0(1, .5) = .5[\beta^2 + \beta(1-\beta)\alpha^2 + \beta(1-\beta)\alpha(1-\alpha)W_0(1, .5)]$$
$$+ .5[\alpha^2 + \alpha(1-\alpha)\beta^2 + \alpha(1-\alpha)\beta(1-\beta)W_0(1, .5)]$$

and, therefore,

$$(3.2) \qquad W_0(1, .5) = \frac{\beta^2 + \alpha^2 + \alpha\beta^2 + \alpha^2\beta - 2\alpha^2\beta^2}{2[1 - \alpha(1-\alpha)\beta(1-\beta)]} .$$

Since $\varphi_{\mathcal{R}}[(1-\alpha)/(2-\alpha-\beta)] = .5$ we have

$$(3.3) \qquad W_0(2, (1-\alpha)/(2-\alpha-\beta))$$
$$= \left[\frac{1-\alpha}{2-\alpha-\beta}\,\beta + \frac{1-\beta}{2-\alpha-\beta}\,\alpha\right] + \left[\frac{1-\alpha}{2-\alpha-\beta}\,(1-\beta) + \frac{1-\beta}{2-\alpha-\beta}\,(1-\alpha)\right]W_0(1, .5) .$$

As an alternative to τ_0 for starting at $k = 2$ with $r = (1-\alpha)/(2-\alpha-\beta)$, consider strategy τ_1: pull $\mathcal{L}$ and thereafter use Ψ_0. Let

$$W_1(2, (1-\alpha)/(2-\alpha-\beta)) = P_{\tau_1}(k_n \to 3) .$$

By virtue of (3.1), when 0 or 3 is not reached before stage 7 the pulls at stages 1 through 6 are $\mathcal{L}$, $\mathcal{R}$, $\mathcal{R}$, $\mathcal{R}$, $\mathcal{R}$, and $\mathcal{R}$, and $r_6 = \varphi_{\mathcal{L}}\sigma_{\mathcal{R}}^3\varphi_{\mathcal{R}}^2 r = r = (1-\alpha)/(2-\alpha-\beta)$. Accordingly,

$$W_1(2,\ (1-\alpha)/(2-\alpha-\beta))$$

$$= \frac{1-\alpha}{2-\alpha-\beta}\,[\alpha + (1-\alpha)\beta^2 + (1-\alpha)\beta^3(1-\beta) + (1-\alpha)\beta^3(1-\beta)^2 W_1(2,\ (1-\alpha)/(2-\alpha-\beta))]$$

$$+ \frac{1-\beta}{2-\alpha-\beta}\,[\beta + (1-\beta)\alpha^2 + (1-\beta)\alpha^3(1-\alpha) + (1-\beta)\alpha^3(1-\alpha)^2 W_1(2,\ (1-\alpha)/(2-\alpha-\beta))]$$

and, therefore,

$$(3.4) \qquad W_1(2,\ (1-\alpha)/(2-\alpha-\beta))$$

$$= \frac{\alpha(1-\alpha) + \beta(1-\beta) + \beta^2(1-\alpha)^2 + \alpha^2(1-\beta)^2 + [(1-\alpha)^2+(1-\beta)^2]C}{[1 - (1-\alpha)(1-\beta)C][2-\alpha-\beta]}\ ,$$

where $C = \alpha^3(1-\alpha) = \beta^3(1-\beta)$.

For $\alpha = 7/15$ and $\beta = 14/15$, which satisfy (3.1), we obtain $(1-\alpha)/(2-\alpha-\beta) = 8/9$,

$$W_1(2,\ 8/9) = \frac{32777675}{34106019} > .9610$$

from (3.4), and

$$W_0(2,\ 8/9) = \frac{6465375}{6728535} < .9609$$

from (3.3) and (3.2). ∎

Numerical calculations for a fine mesh of α's (and their corresponding β's) indicate that the inequality $W_1 > W_0$ obtained for $\alpha = 7/15$, $\beta = 14/15$ holds for any (α,β) satisfying (3.1) when starting at $k = 2$ and $r = (1-\alpha)/(2-\alpha-\beta)$. Even for $\alpha = 7/15$, $\beta = 14/15$, $k_0 = 2$, and $r_0 = 8/9$ in Example 3.1 we do not know that τ_1 is an optimal strategy. We conjecture that $\mathfrak{L}$ is optimal starting

at $k = 2$ and $r = (1-\alpha)/(2-\alpha-\beta)$ -- an advantage of pulling $\mathfrak{L}$ initially is that the decision maker then has a strong opinion about which arm is better when the pull fails and fortune $k = 1$ is reached (in the example, $\varphi_{\mathfrak{L}} r = 64/65$ and $u(1, 14/15) = .9289$).

In the next two sections we calculate the optimal schemes for two special classes of distributions. The myopic scheme Ψ_0 is seen to be optimal for both classes.

4. The case $\alpha = 0$. In this section the smaller success probability is assumed to be 0. We shall exclude some trivial cases, the most interesting of which is $g = -\infty$, $G < \infty$, $.5 \leq \beta < 1$. For this case either arm is optimal at any particular (k,r) with $0 < r < 1$, but an arm that gives all failures cannot be pulled forever, so not all schemes Ψ satisfying $\Psi(k,0) = \mathcal{L}$ and $\Psi(k,1) = \mathcal{R}$ are optimal.

The next theorem says, quite generally, that myopic strategies are optimal when $\alpha = 0$, and that there are no other optimal strategies. The proof depends on Theorem 2.2 of (Berry and Fristedt 1979) and the concept of an excessive function as used there. A function $V(k,r)$ is excessive if, for $\mathcal{G} = \mathcal{L}$ and $\mathcal{G} = \mathcal{R}$,

$$V(k,r) \geq E_{\mathcal{G}}^{(k,r)} V(k_1, r_1)$$

where the subscript denotes the arm pulled at stage 1.

THEOREM 4.1. Suppose $0 = \alpha < \beta < 1$ and either (i) g and G are both finite, (ii) $G < \infty$ and $\beta < .5$, or (iii) $g > -\infty$ and $\infty > .5$. Then a scheme Ψ is optimal for $Q_{\alpha,\beta}$ if and only if

$$\Psi(k,r) = \mathcal{L} \text{ when } r < .5$$
$$= \mathcal{R} \text{ when } r > .5. \tag{4.1}$$

REMARK. This theorem has obvious intuitive appeal, even in the presence of Example 3.1. Nevertheless, the proof we give involves some calculation and goes only part way towards corresponding to intuition.

PROOF OF THEOREM 4.1. One of the schemes that satisfies (4.1) is Ψ_0, defined in (2.2). Define

$$V_0(k,r) = P_{\Psi_0}^{(k,r)}(k_n \to G).$$

Our first step will be to obtain an expression for V_0.

Under Ψ_0, once a success is obtained with an arm that arm is pulled thereafter (since $\sigma_{\mathcal{R}} r = 1$ and $\sigma_{\mathcal{L}} r = 0$). Assume $.5 \leq r < 1$, then Ψ_0 indicates a pull of $\mathcal{R}$ initially, and henceforth, until the probability that $\rho = \beta$ is less than .5. Let $J(r)$ denote the number of consecutive failures on $\mathcal{R}$ until Ψ_0 indicates a pull of $\mathcal{L}$; that is,

$$J(r) = \inf\{j: \ \varphi_{\mathcal{R}}^{j} r < .5\} ,$$

or,

$$\varphi_{\mathcal{R}}^{J(r)} r < .5 \leq \varphi_{\mathcal{R}}^{J(r)-1} r. \tag{4.2}$$

If $\mathcal{L}$ is pulled after $J(r)$ consecutive failures on $\mathcal{R}$ and it also yields failure, then $\mathcal{R}$ is again indicated by Ψ_0; for, in view of (1.2) and (4.2),

$$\varphi_{\mathcal{L}}\varphi_{\mathcal{R}}^{J(r)} r = \varphi_{\mathcal{R}}^{J(r)-1} r \geq .5.$$

Since the process induced by Ψ_0 now steps back and forth between the probabilities of $\rho = \beta$ given as upper and lower limits in (4.2), the entire scheme Ψ_0 is clear when $r \geq .5$: $\mathcal{R}$ is pulled a number, $J(r)$, of times, if all are failures then $\mathcal{L}$ is used, and $\mathcal{R}$ and $\mathcal{L}$ are alternated as long as failures are obtained; as soon as an arm yields

a success it is pulled henceforth.

Therefore,

$$(4.3) \qquad V_0(k,r) = \sum_{j=0}^{k-g-1} q(j,r,\beta)u(k+1-j, \beta), \ .5 \leq r < 1,$$

where, $u(i,\beta)$ is the probability of approaching G starting at i using only a β-arm and has a well-known expression given in Section 1 of Berry and Fristedt (1979), and $q(j,r,\beta)$, the probability under Ψ_0 that $Z_1 = \ldots = Z_j = -1$ and $Z_{j+1} = 1$, satisfies:

$$(4.4) \qquad \begin{aligned} q(j,r,\beta) &= r(1-\beta)^j\beta \ , \ j < J(r) - 1 \\ &= (1-r)(1-\beta)^m\beta \ , \ j = J(r) + 2m, \ m = 0, 1, \ldots \\ &= r(1-\beta)^{J(r)+m}\beta \ , \ j = J(r) + 2m + 1, \ m = 0, 1, \ldots . \end{aligned}$$

Together with the obvious relations,

$$(4.5) \qquad V_0(k,1) = u(k,\beta)$$

and

$$(4.6) \qquad V_0(k,r) = V_0(k, 1-r) \ ,$$

(4.3), (4.4), and (4.2) give V_0 explicitly.

Our next step will be to prove that V_0 is excessive. We do so by showing

$$(4.7) \qquad V_0(k,r) > E_{\mathfrak{L}}^{(k,r)}V_0(k_1, r_1) \text{ if } r > .5 \ ,$$

$$(4.8) \qquad V_0(k,.5) = E_{\mathfrak{L}}^{(k,r)}V_0(k_1, r_1) \ ,$$

$$(4.9) \qquad V_0(k,r) > E_{\mathfrak{R}}^{(k,r)}V_0(k_1, r_1) \text{ if } r < .5 \ .$$

By symmetry ($r \to 1-r$ when the labels on the arms are exchanged) (4.8) is immediate. Also by symmetry, (4.7) and (4.9) are equivalent. Accordingly, we may restrict attention to (4.7).

The strategy implicit in the right side of (4.7), call it τ^*, begins with a pull of $\mathcal{L}$ and then uses Ψ_0. So, following τ^*, if a failure is obtained on the initial pull of $\mathcal{L}$, $\mathcal{R}$ is pulled $J(\varphi_{\mathcal{L}} r) = J(r) + 1$ times and then, as long as failures are obtained, arms $\mathcal{L}$ and $\mathcal{R}$ are pulled alternately; again, an arm that yields a success is pulled indefinitely thereafter. Therefore,

$$(4.10) \qquad E_{\mathcal{L}}^{(k,r)} V_0(k_1, r_1) = P_{\tau^*}^{(k,r)}(k_n \to G)$$
$$= \sum_{j=0}^{k-g-1} q^*(j,r,\beta) u(k+1-j, \beta),$$

where $q^*(j,r,\beta)$ denotes the probability that $Z_1 = \ldots = Z_j = -1$ and $Z_{j+1} = 1$ under Ψ^*. We easily obtain

$$q^*(j,r,\beta) = (1-r)\beta, \qquad j = 0$$
$$= r(1-\beta)^{j-1}\beta, \qquad 1 \le j \le J(r) + 1$$
$$= (1-r)(1-\beta)^m\beta, \qquad j = J(r) + 2m, \; m = 1, 2, \ldots$$
$$= r(1-\beta)^{J(r)+m}, \qquad j = J(r) + 2m + 1, \; m = 1, 2, \ldots.$$

We see that, for $r > .5$,

$$\sum_{i=0}^{j} [q(i,r,\beta) - q^*(k,r,\beta)] \ge 0$$

with strict inequality in case $j = 0$. This fact, combined with the fact

that $u(k+1-j, \beta)$ is a strictly decreasing function of j, implies that $V_0(k,r)$ is larger than (4.10). Hence (4.7) holds and V_0 is excessive.

Since (2.3) of (Berry and Fristedt 1979) is clearly satisfied, Theorem 2.2 of that paper implies that Ψ_0 is optimal for $Q_{\sim 0,\beta}$. Since (4.7) and (4.9) are strict inequalities, the only conserving pulls (c.f. (2.1)) not consistent with Ψ_0 are pulls of $\mathfrak{L}$ when $r = .5$. ∎

5. The case $\alpha + \beta = 1$. In this section we assume that $\alpha = 1 - \beta$. Except for the trivial cases of $\alpha = 0$ and $g = -\infty$, Theorem 5.1 specifies all optimal schemes. An interesting feature of the theorem is the large number of such schemes. Typically, but depending on r, when k is large there is a great deal of flexibility in choosing an arm to pull while behaving optimally. This flexibility is possible because of a special characteristic of posterior distributions when $\alpha = 1 - \beta$; namely,

$$(5.1) \qquad \sigma_{\mathfrak{G}}^{m}\varphi_{\mathfrak{G}}^{m} r = r$$

for all m and $\mathfrak{G} = \mathfrak{R}$ or $\mathfrak{L}$.

The myopic schemes are included in the class shown to contain all optimal schemes in Theorem 5.1. We remark that if the time to reach G were important -- say we were to maximize $P(k_n = G \text{ for some } n \leq N < \infty)$ -- then it would be reasonable to expect only myopic schemes to be optimal.

An interesting feature of the schemes satisfying (5.2) in the theorem is their lack of dependence on G. The proof will show that it is only net failures that lessen the flexibility of optimal schemes. So, it is only proximity to g that affects the class of optimal schemes.

THEOREM 5.1. Suppose $0 < \alpha = 1 - \beta < .5$ and $g > -\infty$. A scheme Ψ is optimal for $Q_{\alpha,\beta}$ if and only if

(5.2) $\Psi(k,r) = \mathcal{R}$ if $r/(1-r) > (\beta/\alpha)^{k-g-1}$

$= \mathcal{L}$ if $(1-r)/r < (\beta/\alpha)^{k-g-1}$.

REMARK. A statement equivalent to (5.2) is

$\Psi(k,r) = \mathcal{R}$ if $r > r^*(k)$

$= \mathcal{L}$ if $r < 1-r^*(k)$,

where

$$r^*(k) = \left[1 + (\alpha/\beta)^{k-g-1}\right]^{-1} .$$

PROOF OF THEOREM 5.1. We may assume $g = 0$. For definiteness we consider a particular scheme given in (5.2), the one which pulls $\mathcal{L}$ whenever such a pull is consistent with (5.2):

$\Psi_1(k,r) = \mathcal{R}$ if $r/(1-r) > (\beta/\alpha)^{k-1}$

$= \mathcal{L}$ otherwise.

Let

$$V_1(k,r) = P_{\Psi_1}^{(k,r)}(k_n \to G).$$

Easy calculations yield:

(5.3) $$\sigma_{\mathcal{R}} r/(1-\sigma_{\mathcal{R}} r) = \varphi_{\mathcal{L}} r/(1-\varphi_{\mathcal{L}} r) = \frac{r}{1-r} \cdot \frac{\beta}{\alpha} ;$$

(5.4) $$\sigma_{\mathcal{L}} r/(1-\sigma_{\mathcal{L}} r) = \varphi_{\mathcal{R}} r/(1-\varphi_{\mathcal{R}} r) = \frac{r}{1-r} \cdot \frac{\alpha}{\beta} .$$

When starting at (k,r), we see from (5.1), (5.3), and (5.4) that the strategy induced by Ψ_1 is as follows: if $r/(1-r) > (\beta/\alpha)^{k-1}$ then $\mathcal{R}$ is pulled at every stage; if $r/(1-r) \leq (\beta/\alpha)^{-k+1}$ then

$\mathcal{L}$ is pulled at every stage; and if $(\beta/\alpha)^{k-2j-1} < r/(1-r) \leq (\beta/\alpha)^{k-2j+1}$ for some $j \in [1, k-1]$ then $\mathcal{L}$ is pulled until (if ever) fortune $k-j$ is reached and $\mathcal{R}$ is pulled thereafter. For completeness, let $j = 0$ and $j = k$ correspond, respectively, to the first two cases listed.

For the upcoming discussion the reader may find it useful to picture the various parallel lines, $(k-2j+1)\log[\beta/\alpha]$, for $j = 1, 2, \ldots$. Pulling $\mathcal{R}$ generates movement of $(k, \log[r/(1-r)])$ parallel to these lines, so that the value of j as defined in Ψ_1 is unchanged by pulls of $\mathcal{R}$. On the other hand, pulling $\mathcal{L}$ generates movement parallel to $(-k+1)\log[\beta/\alpha]$. Exactly one of the above-mentioned family of parallel lines is crossed each time $\mathcal{L}$ is pulled and j is increased or decreased by 1 according as $\mathcal{L}$ yields a success or a failure -- except that $j = 0$ and $j = k$ may not be changed. (This picture also provides a graphical demonstration of (5.1).)

We claim that, for $j = 0, 1, \ldots, k$,

$$(5.5) \qquad V_1(k,r) = r[u(j,\alpha) + u(k-j, \beta)] + (1-r)[u(j,\beta) + u(k-j, \alpha)] ,$$

where, as in Section 4, $u(i,\gamma)$ is the probability of approaching G starting at i using only a γ-arm. The reader is cautioned that (5.5) does not hold for general (α,β).

To derive (5.5) we introduce the goal G into the u notation: $u(k, \gamma; G)$; $g = 0$ is still understood throughout. From the definition

of Ψ_1 we have, for $j = 0, 1, \ldots, k$,

$$
\begin{aligned}
(5.6) \qquad V_1(k,r) &= r\{u(j, \alpha; G-k+j) + [1-u(j, \alpha; G-k+j)]u(k-j, \beta; G)\} \\
&\quad + (1-r)\{u(j, \beta; G-k+j) + [1-u(j, \beta; G-k+j)]u(k-j, \alpha; G)\} \\
&= r\{u(j, \alpha; G-k+j)[1-u(k-j, \beta; G)] + u(k-j, \beta; G)\} \\
&\quad + (1-r)\{u(j, \beta; G-k+j)[1-u(k-j, \alpha; G)] + u(k-j, \alpha; G)\}.
\end{aligned}
$$

Equation (5.5) now follows from (5.6) using

$$1-u(k-j, 1-\gamma; G) = u(G-k+j, \gamma; G)$$

and

$$u(j, \gamma; G-k+j)u(G-k+j, \gamma; G) = u(j, \gamma; G).$$

The next step is to show that V_1 is excessive. We require

$$(5.7) \qquad V_1(k,r) > E_{\mathfrak{L}}^{(k,r)} V_1(k_1,r_1) \quad \text{if} \quad r/(1-r) > (\beta/\alpha)^{k-1}$$

$$(5.8) \qquad V_1(k,r) > E_{\mathfrak{R}}^{(k,r)} V_1(k_1,r_1) \quad \text{if} \quad r/(1-r) < (\beta/\alpha)^{-k+1}$$

$$(5.9) \qquad V_1(k,r) = E_{\mathfrak{R}}^{(k,r)} V_1(k_1,r_1) \quad \text{if} \quad (\beta/\alpha)^{-k+1} \leq r/(1-r) \leq (\beta/\alpha)^{k-1}.$$

We will show that (5.7) holds and omit the proofs of (5.8) and (5.9); their demonstration is similar to that of (5.7) but is somewhat easier since, as mentioned previously, the value of j in Ψ_1 is not changed by pulling $\mathfrak{R}$.

When $r/(1-r) > (\beta/\alpha)^{k-1}$ and $\mathfrak{L}$ is pulled yielding a failure, then j remains equal to 0. When $\mathfrak{L}$ yields a success, then j may remain at 0 or increase to 1. Accordingly, we consider two cases.

Assume first that $r/(1-r) > (\beta/\alpha)^{k+1}$. Then, since j remains at 0 for both success and failure on $\mathcal{L}$, (5.5) yields:

$$E_{\mathcal{L}}^{(k,r)}V_1(k_1,r_1) = [r\alpha + (1-r)\beta]V_1(k+1, \sigma_{\mathcal{L}}r) + [r\beta + (1-r)\alpha]V_1(k-1, \varphi_{\mathcal{L}}r)$$

$$= r\alpha u(k+1, \beta) + (1-r)\beta u(k+1, \alpha) + r\beta u(k-1, \beta) + (1-r)\alpha u(k-1, \alpha)$$

$$= ru(k,\beta) + (1-r)u(k,\alpha) - (\beta-\alpha)[ru(k+1, \beta) - (1-r)u(k+1, \alpha) - ru(k-1, \beta) + (1-r)u(k-1, \alpha)] .$$

When $G < \infty$,

$$E_{\mathcal{L}}^{(k,r)}V_1(k_1, r_1) - V_1(k,r) =$$

$$= -(\beta-\alpha)\frac{1 - \left(\frac{\beta}{\alpha}\right)^2}{1 - \left(\frac{\beta}{\alpha}\right)^G}\left[r\left(\frac{\beta}{\alpha}\right)^{G-k-1} - (1-r)\left(\frac{\beta}{\alpha}\right)^{k-1}\right]$$

$$< -(\beta-\alpha)\frac{1 - \left(\frac{\beta}{\alpha}\right)^2}{1 - \left(\frac{\beta}{\alpha}\right)^G}(1-r)\left[\left(\frac{\beta}{\alpha}\right)^G - \left(\frac{\beta}{\alpha}\right)^{k-1}\right] < 0 ;$$

$G = \infty$ is similar.

Now assume that $(\beta/\alpha)^{k-1} < r/(1-r) \leq (\beta/\alpha)^{k+1}$, then j remains at 0 if $\mathcal{L}$ yields a failure and becomes 1 after a success; for,

$$\left(\frac{\beta}{\alpha}\right)^{k-2} < \frac{\sigma_{\mathcal{L}}r}{1 - \sigma_{\mathcal{L}}r} \leq \left(\frac{\beta}{\alpha}\right)^{k} \Leftrightarrow \left(\frac{\beta}{\alpha}\right)^{k_1-3} < \frac{\sigma_{\mathcal{L}}r}{1 - \sigma_{\mathcal{L}}r} \leq \left(\frac{\beta}{\alpha}\right)^{k_1-1} .$$

We have,

$$E_{\mathcal{L}}^{(k,r)}V_1(k_1,r_1) = (r\alpha + (1-r)\beta)V_1\left(k+1, \frac{r\alpha}{r\alpha + (1-r)\beta}\right)$$

$$+ (r\beta + (1-r)\alpha)V_1\left(k-1, \frac{r\beta}{r\beta + (1-r)\alpha}\right)$$

$$= r\alpha[u(1,\alpha) + u(k,\beta)] + (1-r)\beta[u(1,\beta) + u(k,\alpha)]$$

$$+ r\beta[u(0,\alpha) + u(k-1, \beta)] + (1-r)\alpha[u(0,\beta) + u(k-1, \alpha)] \quad ,$$

according to (5.5). When $G < \infty$,

(5.10) $$E_{\mathcal{L}}^{(k,r)}V_1(k_1,r_1) - V_1(k,r)$$

$$= r\{\alpha u(1,\alpha) + \beta[u(k-1, \beta) - u(k,\beta)]\}$$

$$+ (1-r)\{\beta u(1,\beta) + \alpha[u(k-1, \alpha) - u(k,\alpha)]\}$$

$$= r\left[\frac{\alpha - \beta}{1 - \left(\frac{\beta}{\alpha}\right)^G} + \frac{(\alpha-\beta)\left(\frac{\alpha}{\beta}\right)^{k-1}}{1 - \left(\frac{\alpha}{\beta}\right)^G}\right]$$

$$+ (1-r)\left[\frac{\beta - \alpha}{1 - \left(\frac{\alpha}{\beta}\right)^G} + \frac{(\beta-\alpha)\left(\frac{\beta}{\alpha}\right)^{k-1}}{1 - \left(\frac{\beta}{\alpha}\right)^G}\right]$$

$$= (\beta-\alpha)\frac{1 - \left(\frac{\beta}{\alpha}\right)^{G-k+1}}{1 - \left(\frac{\beta}{\alpha}\right)^G}\left[(1-r)\left(\frac{\beta}{\alpha}\right)^{k-1} - r\right]$$

$$< 0,$$

since $\alpha < \beta$ and $r/(1-r) > (\beta/\alpha)^{k-1}$; $G = \infty$ is similar. Notice that the difference in (5.10) is small when (k,r) is barely in the "$j = 0$ region."

It can be shown from (5.5) that $V_1(k,r) \to 1$ uniformly in r as

$k \to G$; therefore, we can apply Theorem 2.2 of the companion paper (Fristedt and Berry 1979) to conclude that Ψ_1 is optimal. From (5.7), (5.8), and (5.9) we see that the strategies satisfying (5.2) are exactly the conserving strategies and, therefore, according to Theorem 2.1 of Fristedt and Berry (1979), are exactly the optimal strategies. ∎

Set $g = 0$ and fix $j \in [1, k-1]$. Consider r satisfying

$$(5.11) \qquad (\beta/\alpha)^{k-2j-1} < r/(1-r) \leq (\beta/\alpha)^{k-2j+1}.$$

According to the proof of Theorem 5.1 the probabilities of approaching G rather than reaching $g = 0$ are in fact identical for all strategies satisfying: never pull $\mathfrak{L}$ after j more failures than successes have been obtained on $\mathfrak{L}$ and never pull $\mathfrak{R}$ after $k-j$ more failures than successes have been obtained on $\mathfrak{R}$. Since the probabilities of approaching G depend linearly on r and are equal for all such strategies for r belonging to the entire interval defined by (5.11), they are also equal for $r = 0$ and $r = 1$. Thus we have the following corollary about arms for which the probabilities of success are known.

COROLLARY 5.1. Let i, j, and m be positive integers and let $\gamma \in (0,1)$. All past-history-dependent sequences of pulls of a γ-arm and a $(1-\gamma)$-arm (calling heads vs. tails in biased coin-tossing, for example) satisfying the condition: the γ-arm is never pulled after i more failures than successes have been obtained with it

and the $(1-\gamma)$-arm is never pulled after j more failures than successes have been obtained with it, all have the same probability of eventually yielding a total excess of m successes over failures.

6. An extension. The class of distributions $\underset{\sim}{Q}$ is quite special and, as mentioned in Section 1, there is little hope of handling a much larger class. Still, the results of Sections 4 and 5 can be seen to apply to a somewhat larger class by a simple observation: For an arbitrary distribution on the unit square there is no loss by conditioning on $\lambda \neq \rho$. A corollary is that nothing is lost by moving some mass to the line $\lambda = \rho$ so long as ratios of other probabilities are not changed. So optimal schemes for any $\underset{\sim}{Q}_{\alpha,\beta}$ are also optimal for

$$p\underset{\sim}{Q}_{\alpha,\beta} + (1-p)T,$$

where T is a distribution measure on $\{(\lambda,\rho): \lambda = \rho\}$ and $0 < p < 1$. In particular, Theorems 4.1 and 5.1 apply with evident modifications.

As an example, suppose mass is moved from $\{(\alpha,\beta), (\beta,\alpha)\}$ to $\{(\alpha,\alpha), (\beta,\beta)\}$ so that the new probabilities of (α,β) and (β,α) are in the same ratio, $r/(1-r)$, and λ and ρ are independent. Since Theorems 4.1 and 5.1 apply for all r, it is optimal to always pull the arm with the higher probability of being the β-arm whenever λ and ρ are independent and both have measures on $\{\alpha,\beta\}$, with $\alpha = 0$ or $\alpha = 1-\beta<.5$. Berry (1972, Section 8) makes an analogous extension for the classical two-armed bandit.

References

Berry, Donald A. (1972). A Bernoulli two-armed bandit. Annals of Mathematical Statistics 43 871-897.

Berry, Donald A. and Bert Fristedt (1979). Two armed bandits with a goal; I: one arm known, University of Minnesota, School of Statistics Technical Report No. 344, August, 1979. Submitted to Advances in Applied Probability.

Fabius, J. and van Zwet, W.R. (1970). Some remarks on the two-armed bandit. Annals of Mathematical Statistics 41 1906-1916.

Feldman, Dorian (1962). Contributions to the "two-armed bandit" problem. Annals of Mathematical Statistics 33 847-856.

Kelley, Thomas A. (1974). A note on the Bernoulli two-armed bandit. Annals of Statistics 2 1056-1062.